Hindawi
Complexity
Volume 2021, Article ID 6117513, 11 pages
https://doi.org/10.1155/2021/6117513

Research Article

Comparing the Forecast Performance of Advanced Statistical and Machine Learning Techniques Using Huge Big Data: Evidence from Monte Carlo Experiments

Faridoon Khan,[1] Amena Urooj,[1] Saud Ahmed Khan,[1] Abdelaziz Alsubie,[2] Zahra Almaspoor ,[3] and Sara Muhammadullah[1]

[1]*PIDE School of Economics, Pakistan Institute of Development Economics, Islamabad, Pakistan*
[2]*Department of Basic Sciences, College of Science and Theoretical Studies, Saudi Electronic University, Riyadh, Saudi Arabia*
[3]*Department of Statistics, Yazd University, Yazd 89175-741, Iran*

Correspondence should be addressed to Zahra Almaspoor; z.almaspoor@stu.yazd.ac.ir

Received 12 October 2021; Revised 17 November 2021; Accepted 30 November 2021; Published 14 December 2021

Academic Editor: Paulo Jorge Silveira Ferreira

This research compares factor models based on principal component analysis (PCA) and partial least squares (PLS) with Autometrics, elastic smoothly clipped absolute deviation (E-SCAD), and minimax concave penalty (MCP) under different simulated schemes like multicollinearity, heteroscedasticity, and autocorrelation. The comparison is made with varying sample size and covariates. We found that in the presence of low and moderate multicollinearity, MCP often produces superior forecasts in contrast to small sample case, whereas E-SCAD remains better. In the case of high multicollinearity, the PLS-based factor model remained dominant, but asymptotically the prediction accuracy of E-SCAD significantly enhances compared to other methods. Under heteroscedasticity, MCP performs very well and most of the time beats the rival methods. In some circumstances under large samples, Autometrics provides a similar forecast as MCP. In the presence of low and moderate autocorrelation, MCP shows outstanding forecasting performance except for the small sample case, whereas E-SCAD produces a remarkable forecast. In the case of extreme autocorrelation, E-SCAD outperforms the rival techniques under both the small and medium samples, but further augmentation in sample size enables MCP forecast more accurate comparatively. To compare the predictive ability of all methods, we split the data into two halves (i.e., data over 1973–2007 as training data and data over 2008–2020 as testing data). Based on the root mean square error and mean absolute error, the PLS-based factor model outperforms the competitor models in terms of forecasting performance.

1. Introduction

The prediction of macroeconomic variables is very important under macroeconomic studies, monetary policy analysis, and environmental economics. Accurate forecasts induce sound insights into mechanisms of dynamic economies [1], more effective monetary policies [2], and better portfolio management and hedging strategies [3]. In the data-rich environment existing these days, many macroeconomic series are tracked by economists and decision-makers.

Low-dimensional models often include some prespecified economic covariates for instance vector autoregression and therefore have a complication in capturing the dynamic and complex patterns, which contain huge panels of time series [4]. It is a fact that missing important variable(s) leads to an underspecified model, inducing biased results. There is an intense need to propose updated statistical models and analysis frameworks with the purpose of expanding the low-dimensional counterparts for improved forecasts. Thus, in the recent era, the analysis of "Big Data" has become the core of economics research. This in turn has resulted in special attention being paid to the huge class of techniques that are available in the domain of machine learning, dimension reduction, and penalized

regression [5, 6]. Recently, in the regression context, Doornik and Hendry [7] categorized Big Data into three classes: tall big data, huge big data, and fat big data. Each type can be defined as follows:

(i) Tall big data: more observations and several covariates ($N >> P$)

(ii) Huge big data: more observations and more covariates ($N > P$)

(iii) Fat big data: fewer observations and more covariates ($N < P$)

where N and P represent the number of observations and covariates, respectively. We graphically represent the Big Data in Figure 1.

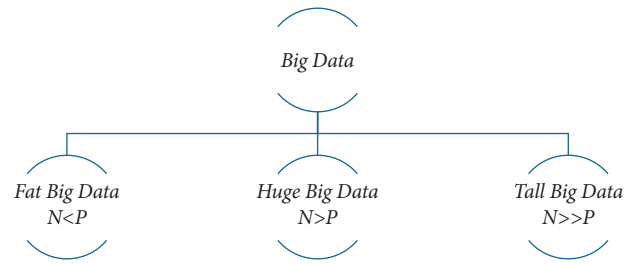

Figure 1: Schematic representation of big data.

There are many related studies on macroeconomic forecasting based on factor models and machine learning techniques. In the last two decades, forecasting studies using large-scale datasets and pseudo-out-of-sample forecast incorporate those by Artis et al. [8]; Boivin and Ng [9, 10]; Forni et al. [11]; Armah and Swanson [12, 13]; Stock and Watson [14–18]; Varian [5]; Kim and Swanson [19, 20]; Castle et al. [21, 22]; Luciani [23]; Kristensen [24]; Swanson and Xiong [6, 25]; Tu and Lee [26]; Swanson et al. [27]; Maehashi and Shintani [28]; Kim and Ko [29]; Kim et al. [30]; Abdić et al. [31]; and and Kim and Shi [32].

Moreover, Stock and Watson [17] elaborately discussed the past studies on the utility of factor models forecasting. There is an intensive and growing body of literature in this area. Few of them are relevant, as they address both theoretical and empirical problems, including Armah and Swanson [12, 13]; Artis et al. [8]; Bai and Ng [1, 33, 34], Banerjee and Marcellino [35]; Boivin and Ng [9, 10], Ding and Hwang [36]; Dufour and Stevanovic [37]; Stock and Watson [15–18]; and Smeekes and Wijler [38].

The abovementioned papers consider principal component analysis, independent component analysis, and sparse principal component analysis for the construction of the factor model. However, there is also a small and growing body of literature investigating the classical approach (Autometrics) in the context of macroeconomic forecasting [7, 21, 22]. We failed to discover any paper to date that has investigated the use of partial least squares (PLS) theoretically in our context. However, the method has been applied empirically in various fields. Apart from this, some papers have utilized shrinkage methods like ridge regression, lasso, elastic net, adaptive lasso, and nonnegative garrote, but none of the papers to date have used the updated forms of shrinkage methods in our context.

Filling the gaps, this work implements some updated techniques of big data to increment literature of macroeconomic forecasting theoretically as well as empirically. From the dimension reduction aspect, we build factor models intending to highlight the importance of such models for macroeconomic prediction. Particularly, while building factor models, we employ principal component analysis (PCA) and partial least squares (PLS). In addition, we also assess the last version of the classical approach (Autometrics) and the updated version of shrinkage methods including elastic smoothly clipped absolute deviation (E-SCAD) and minimax concave penalty (MCP). We evaluate the performance of these techniques in a simulation setting where the true data generating process (DGP) of the factor model is used. To summarize the whole discussion, our prime contribution comes in the form of comparison of updated shrinkage methods and Autometrics with factor models through forecasting under the simulated scenarios having multicollinearity, heteroscedasticity, and autocorrelation along with application to macroeconomic data to provide a conclusive solution to predictability. The study aims to produce an improved method to help policymakers; the improved tool is not restricted to workers' remittances or the stock market (*in our case*) but is valid for any time series.

The remaining part of the paper is organized as follows. In Section 2, we provide a detailed discussion regarding factor models based on principal component analysis and partial least squares. In Section 3, we discuss big data techniques, such as the classical approach and shrinkage methods. Monte Carlo evidence on the comparative performance of several forecasting techniques is discussed in Section 4. Empirical findings are given in Section 5. Section 6 provides concluding remarks.

2. Methods

The techniques we intend to apply in subsequent sections are reported in Figure 2.

This study aims to compare the predictive ability of factor models based on principal component analysis and partial least squares with Autometrics, elastic smoothly clipped absolute deviation (E-SCAD), and minimax concave penalty under different scenarios like multicollinearity, heteroscedasticity, and autocorrelation. Macroeconomic and financial datasets are used for the analysis of the real phenomenon.

2.1. Factor Models. The notion of factor models also called diffusion index entails the utility of properly extracted hidden common factors that have been distilled from a huge set of features as inputs in the identification of the parsimonious models. To be more specific, let X be an $N \times P$ dimensional matrix of data points and define $N \times k$ dimensional matrix of latent factors.

Stock and Watson [17] have delineated in depth the literature regarding forecasting through factor models. In

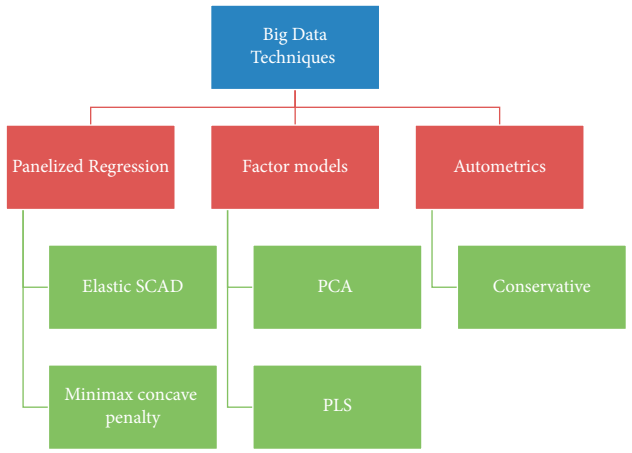

Figure 2: Methods of big data.

the below detailed discussion of factor model methodology, we follow Stock and Watson [15]:

$$X = F\varphi' + \varepsilon, \tag{1}$$

where ε represent the random error matrix, φ' is the $P \times k$ coefficients matrix, and F is a factor matrix of $N \times k$ dimension.

We construct the following forecasting model based on the work of Bai and Ng [39], Kim and Swanson [19], and Stock and Watson [15]:

$$Y_{t+h} = F_t \gamma_F + e_{t+h}, \tag{2}$$

where Y_{t+h} is an outcome variable to be forecasted, h shows the forecast horizon, and F_t is the vector of factors with a dimension, distilled from F in equation (1). The associated coefficient γ_F is a vector of unknown parameters, and e_{t+h} is the random error. The whole process of factor model forecasting consists of two steps: in the first step, we estimate k latent (unobserved) factors, represented by $\widehat{F}$, from P observable predictors. To gain convenient dimension reduction, k is supposed to be much smaller than P (i.e., $k \ll P$). In the second step, we estimate $\widehat{\gamma}_F$, by utilizing data at hand with Y_t and $\widehat{F}_t$. Subsequently, an out-of-sample forecast is constructed.

Kim and Swanson [19] utilized the PCA approach to achieve estimates of the unobserved factors, known as principal components (PCs). The latent PCs are uncorrelated which are obtained by using the data projection in the direction of maximal variance, and naturally, the PCs are ordered based on their variance contributions. The first PC reflects the direction of the maximal variance in the data, the second PC reflects the direction that explains the maximal variance in the rest of the orthogonal subspace, and so on.

This approach is most frequently used in the literature of factor analysis because PCs are easily derived via the use of singular value decompositions [15, 33, 34].

Boivin and Ng [10], however, argued that the performance of the factor model is more likely to be worse in prediction if the incorporated factors are dominated by excluding factors. Similarly, Tu and Lee [26] stated that PCA imposes only the factor structure for X and does not consider the outcome variable. It indicates that PCA ignores the dependent variable while performing it. By dint of neglecting the outcome variable at the time of factors, extraction induces an inefficient forecast of the outcome variable. The solution to this problem is given in the next section.

2.2. The Partial Least Squares (PLS) Method. This study looks at another method that is known as partial least squares (PLS) regression developed by Wold [40]. This method is appropriate in a data-rich environment and may be considered as an alternative to PCA-based factor models. Unlike the PCA method, the PLS identifies new factors in a supervised way; that is, it makes use of the response variable to identify new factors that not only approximate the old factors well but are also related to the response variable. Roughly speaking, the PLS approach attempts to find the directions of maximum variance that help in explaining both the response variable and explanatory variables. The PLS for an outcome variable is motivated by a statistical model as follows:

$$Y_t = x_t \gamma_P + e_t, \tag{3}$$

where $x_t = [x_{1,t}, x_{2,t}, \ldots, x_{n,t}]'$ is an $n \times 1$ vector of covariates at time $t = 1, \ldots, T$, γ_P is an $n \times 1$ vector of associated coefficients, and e_t is the disturbance term. Kim and Ko [29] argued that PLS models are useful especially when there are a large number of covariates. Instead of using a model given in (3), one may adopt another data dimension reduction approach through the following linear regression with $Z \times 1$ vector of components $s_t = [s_{1,t}, s_{2,t}, \ldots, s_{Z,t}]$ as follows:

$$\begin{aligned} Y_t &= x_t w \tau + e_t, \\ Y_t &= s_t \tau = e_t. \end{aligned} \tag{4}$$

We define s_t:

$$s_t = w' x_t, \tag{5}$$

where $w = [x_1, x_2, \ldots, x_Z]$ is the $n \times Z$ matrix of each column, $w_z = [w_{1,z}, w_{2,z}, \ldots, w_{n,z}]$, $z = 1, 2, \ldots, Z$, denote the vector of weights on covariates for z factors or components, and τ is the $Z \times 1$ vector of PLS coefficients. We may use the following equation for predicting the k steps ahead model; that is, $\widehat{y}_{t+k}$, $k = 1, 2, \ldots, m$.

$$\widehat{y}_{t+k} = \widehat{\gamma}_k' x_t. \tag{6}$$

3. Classical Approach and Shrinkage Methods

The fundamental comparison of interest here is between automatic selection over variables as against PC and PLS-based factors in terms of prediction. Factors are often regarded as essential to summarize a large amount of information, but the classical approach and shrinkage methods are alternatives.

3.1. Classical Approach. Autometrics is a well-known big data algorithm, which consists of five steps. In the first step,

we begin the process with the construction of a linear model, which refers to the General Unrestricted Model (GUM); in the second step, we obtain the estimates for unknown parameters and test them statistically; the third step entails presearch process; step four delivers the tree-path search; and the last step leads to a selection of the final model.

Doornik [41] elaborately delineated the complete algorithm. The key notion is to commence modeling with a linear model that incorporates all candidate features (GUM). Estimate the GUM by the least squares method and then carry out the statistical tests to validate the congruency of the model. If the estimated GUM contains statistically insignificant coefficients at prespecified criteria, then again estimate the simpler models by utilizing different paths search and ratified by diagnostic tests. As some terminal models are detected, Autometrics undertakes their union testing. The rejected models are discarded, and the union of those terminal models who survived leads to a new GUM for another tree-path search iteration. The whole inspection process proceeds, and the terminal models are statistically checked against their union. If two or more terminal models clear the encompassing tests, then the preselected information criterion decides about the final choice.

The econometric models are achieved by applying Autometrics on the GUM:

$$y_t = \theta_0 + \sum_{u=1}^{m} \sum_{v=0}^{k} \theta_{u,v} x_{u,t-v} + \mu_t. \tag{7}$$

Under Autometrics, two main strategies are commonly used for model selection, a conservative and a superconservative also called Liberal strategy. Our study implements the Liberal strategy, which is typically based on a one percent significance level rather than five percent. In other words, the statistical significance of each estimated coefficient is based on one percent level of significance.

3.2. Shrinkage Methods. An alternative prominent approach to deal with many features is the family of panelized regression methods, which comprises of many techniques, but our study adopts the following updated forms: elastic smoothly clipped absolute deviation and minimax concave penalty.

3.2.1. Elastic Smoothly Clipped Absolute Deviation. Fan and Li [42] added a new penalization technique to literature known as SCAD. The technique is nonconvex and enjoys an oracle property: sparsity, continuity, and unbiasedness. This technique selects useful covariates with their magnitudes asymptotically in an efficient way if the underlying true model is known (i.e., the oracle properties). The SCAD function covers all the limitations faced by the existing methods like ridge and lasso. The penalty function of SCAD is defined as follows:

$$p_k(|\tau|) = k\left\{ I(\tau \le k) + \frac{(\gamma k - \tau)}{(\gamma - 1)k} + I(\tau > k) \right\}. \tag{8}$$

The unknown tuning parameter k was determined by the generalized cross-validation approach, and they assumed the value of γ is 3.7. As given above, the penalty function is continuous, and the resulting solution is given by

$$p_k(|\tau|) = \left\{ \begin{array}{cc} k|\tau| & |\tau| < k \\ -\left(\tau^2 - 2\gamma k|\tau| + k^2\right)/2(\gamma - 1) & k < |\tau| \le \gamma k \\ (\gamma + 1)k^2/2 & |\tau| > \gamma k \end{array} \right\}. \tag{9}$$

The tuning parameters can be induced from the data-driven technique. The limitation of SCAD is that it selects only one variable from a correlated set of predictors. Zeng and Xie [43] extended the SCAD by augmenting L_2 penalty and called it elastic SCAD (E-SCAD). Mathematically, it can be written as

$$\text{pen}_k(|\tau|) = \sum_{d=1}^{D} p_k(|\tau|) + \lambda_{2p} \sum_{d=1}^{m} \alpha_d^2. \tag{10}$$

Due to L_2 penalty, the E-SCAD achieves an additional property along with oracle properties; that is, the penalty function should spur highly correlated features to be in or out of the model simultaneously. Hence, the proposed form selects the whole group of correlated predictors rather than one variable.

3.2.2. Minimax Concave Penalty. Zhang [44] proposed a minimax concave penalty (MCP), which yields the convexity of the penalized loss in sparse regions considerably given specific thresholds for features selection as well as unbiasedness. The MCP is described as follows:

$$S_{\text{MCP}}(t; k) = \left\{ \begin{array}{ll} \text{kt} - \dfrac{t^2}{2\gamma} & \text{if } |t| \le \gamma k \\ \dfrac{1}{2}\gamma k^2 & \text{if } |t| > \gamma k \end{array} \right\}. \tag{11}$$

The tuning parameter ($\gamma > 0$) diminishes the maximum concavity under the following restrictions like unbiasedness and selection of features:

$$\begin{aligned} &\rho(t; k) = 0 \quad \forall\, t \ge \gamma k \quad \rho(0+; k) = k, \\ &\sum_{d=1}^{m} p_d\left(|\alpha_d|; k; \gamma\right). \end{aligned} \tag{12}$$

The dual-tuning parameters in concave penalty regression play a key role in terms of controlling the amount of regularization. Likewise, the concavity of the MCP penalty considerably evades the sparse convexity by dint of diminishing the maximal concavity. In 2010, the author showed that a rise in regularization parameter value leads to bearing more convexity and achieves an almost unbiased penalty. The penalty function of MCP typically belongs to the quadratic spline function.

4. Monte Carlo Evidence on Forecasting Performance

Our simulation part consists of three main scenarios, namely, simulations on a data generating process (DGP) with (i) multicollinearity, (ii) heteroscedasticity, and (iii) autocorrelation. In each simulated scenario, varying the DGP attributes in terms of correlation strength among features, the magnitude of the variance of the error term, and the magnitude of correlation of error term with previous values (lag).

4.1. Data Generating Process. We generate data from the following equation:

$$Y = X_i\beta + \mu. \tag{13}$$

The set of predictors $X_1, X_2, \ldots, X_P$ are generated from multivariate normal distribution as $X_i \sim N(0, \Sigma)$. The same data generating process (DGP) was used by [38] as mentioned in (13) for artificial data generation. Our study considers three types of sample sizes for the simulation experiments. We suppose a dual set of features with altering the number of active (p) and inactive features (q), respectively, as portrayed in Figure 3.

In our simulation experiments, we assume three scenarios as follows: in the first scenario: we generate the pairwise correlation between the predictors (i.e., x_m and x_n as $\operatorname{cov}(x_m, x_n) = \sum^{|m-n|}$). The population covariance matrix is produced in the following way:

$$\sum_P = \begin{bmatrix} 1 & \cdots & \sum^{|n-m|} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \sum^{|m-n|} & \cdots & 1 \end{bmatrix}. \tag{14}$$

While altering the parameter Σ, we obtain different correlation structures. In our work, we assume values for $\Sigma \in \{0.25, 0.5, 0.9\}$ as followed by Xiao and Xu [45]. In the second scenario, we generate the correlation between current and residuals lag (autocorrelation) and symbolized by ρ. The autocorrelation is generated as follows:

$$\mu_t = \rho\epsilon_{t-1} + \varepsilon_t. \tag{15}$$

Our experiments assume the low, moderate, and high cases of autocorrelation, such as $\rho \in \{0.25, 0.5, 0.9\}$. The third scenario is for examining heteroscedasticity (i.e., means that the variance of the error term is not constant and alters across data points by σ_k).

$$E\left(\mu_t^2\right) = \sigma_k. \tag{16}$$

So, we split the variance σ_k into two components (i.e., σ_1 and σ_2). Let us have "n" observations; we set the variance of $(n/2)$ observations as σ_1 and the variance of remaining observations as σ_2. Our simulation experiments assume three cases of heteroscedasticity and set the values of $\pi_i = (\sigma_1/\sigma_2)$, where $i = 1, 2, 3$ as $\pi_i \in \{0.1/0.3, 0.2/0.6, 0.3/0.9\}$. Tenfold cross-validation is executed to determine the optimal value of the tuning parameter(s).

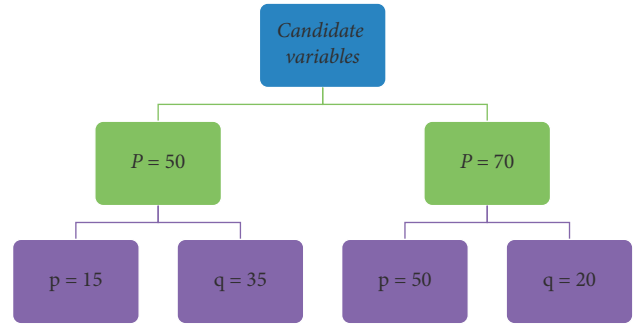

Figure 3: Distribution of candidate variables into relevant and irrelevant variables.

To evaluate the forecasting performance of all methods, we divide each realization such that 80 percent of the data are used to train the models and the remaining data are utilized for models' evaluation followed by [46]. The entire process will be replicated $M = 1000$ times. The average of root mean square (RMSE) and mean absolute error (MAE) are computed over "M" to assess the forecast performance. The smaller the values of RMSE and MAE, the closer the predicted values to the actual values and the better the forecast relatively. For analysis, we have relied on several packages like gets, glmnet, ncvreg, pls, caret, forecast, and Metrics under R programming language.

4.2. Simulation Results. The forecast comparison results derived from Monte Carlo experiments are presented in Tables 1–3. All methods are improving their performance by augmenting the number of observations. Increasing the number of irrelevant and candidate variables adversely affects the predictive ability.

Scenario 1. In the presence of low and moderate multicollinearity, the performance of MCP is superior to other rival methods except for the case of a small sample, where E-SCAD and PLS-based factor models are dominant. To be more specific, in the presence of low and moderate multicollinearity, E-SCAD often produced better forecasts. As we consider the case of high multicollinearity, the PLS-based factor model is superior in particular, while asymptotically E-SCAD outperformed the other methods.

Scenario 2. In the presence of all schemes of heteroscedasticity, the performance of MCP is often better than all competitor models. When the number of predictors is equal to 50, Autometrics provides a similar forecast as MCP in large samples.

Scenario 3. In the presence of low and moderate autocorrelation, the MCP showed an outstanding performance in terms of forecasting particularly when we increase the sample size. In contrast, when $n = 100$, the E-SCAD produced a remarkable forecast. In the case of

Table 1: Forecast comparison under multicollinearity from Monte Carlo simulation (Scenario 1).

Models	$\rho = 0.25,\ P = 50$		$\rho = 0.25,\ P = 70$	
$n = 100/200/400$	RMSE	MAE	RMSE	MAE
MCP	**1.123/1.055/1.027**	**0.908/0.848/0.821**	1.205/**1.069/1.031**	0.971/**0.858/0.825**
E-SCAD	1.135/1.066/1.034	0.917/0.856/0.827	**1.195**/1.086/1.040	**0.961**/0.872/0.831
Autometrics	1.316/1.091/1.027	1.065/0.874/0.822	1.316/1.091/1.042	1.065/0.874/0.834
FM_PCA	3.517/3.210/2.829	2.839/2.576/2.260	4.493/4.305/3.966	3.623/3.458/3.173
FM_PLS	1.528/1.200/1.090	1.235/0.963/0.871	1.921/1.321/1.126	1.551/1.059/0.901
$n = 100/200/400$	$\rho = 0.5,\ P = 50$		$\rho = 0.5,\ P = 70$	
MCP	1.145/**1.056/1.027**	0.925/**0.848/0.821**	1.318/**1.069/1.032**	1.062/**0.858/0.825**
E-SCAD	**1.112**/1.058/1.030	**0.898**/0.849/0.824	**1.168**/1.074/1.035	**0.940**/0.862/0.827
Autometrics	1.156/1.062/1.027	0.931/0.853/0.821	1.473/1.091/1.041	1.191/0.874/0.833
FM_PCA	2.583/2.053/1.705	2.088/1.644/1.365	3.933/3.334/2.700	3.174/2.677/2.164
FM_PLS	1.368/1.161/1.080	1.105/0.932/0.864	1.595/1.248/1.108	1.287/1.001/0.886
$n = 100/200/400$	$\rho = 0.9,\ P = 50$		$\rho = 0.9,\ P = 70$	
MCP	1.484/1.157/1.042	1.198/0.930/0.832	1.764/1.261/1.058	1.424/1.013/0.846
E-SCAD	1.201/**1.060/1.019**	0.968/**0.851/0.814**	1.291/**1.080/1.021**	1.040/**0.867/0.817**
Autometrics	4.363/1.795/1.031	3.528/1.443/0.825	6.589/2.501/1.053	5.333/2.006/0.843
FM_PCA	1.169/1.099/1.075	0.943/0.883/0.859	1.318/1.212/1.165	1.065/0.974/0.932
FM_PLS	**1.138**/1.078/1.043	**0.919**/0.865/0.834	**1.184**/1.095/1.053	**0.959**/0.880/0.842

Note. Bold values indicate a better forecast.

Table 2: Forecast comparison under heteroscedasticity from Monte Carlo simulation (Scenario 2).

Models	$\sigma = 0.1/0.3,\ P = 50$		$\sigma = 0.1/0.3,\ P = 70$	
$n = 100/200/400$	RMSE	MAE	RMSE	MAE
MCP	**0.313/0.306/0.303**	**0.253/0.246/0.242**	**0.321/0.307/0.303**	**0.260/0.246/0.242**
E-SCAD	0.319/0.309/0.304	0.258/0.248/0.243	0.331/0.311/0.305	0.267/0.249/0.243
Autometrics	0.318/0.308/**0.303**	0.256/0.248/**0.242**	0.339/0.313/0.305	0.274/0.250/0.244
FM_PCA	3.373/3.055/2.648	2.723/2.452/2.115	4.382/4.197/3.847	3.534/3.374/3.078
FM_PLS	0.399/0.327/0.311	0.322/0.262/0.249	0.625/0.347/0.317	0.504/0.278/0.253
$n = 100/200/400$	$\sigma = 0.2/0.6,\ P = 50$		$\sigma = 0.2/0.6,\ P = 70$	
MCP	**0.627/0.613/0.606**	**0.507/0.492/0.484**	**0.643/0.614/0.607**	**0.520/0.492/0.485**
E-SCAD	0.637/0.617/0.609	0.515/0.496/0.486	0.659/0.621/0.609	0.532/0.498/0.487
Autometrics	0.636/0.617/**0.606**	0.512/0.496/**0.484**	0.667/0.625/0.610	0.548/0.501/0.488
FM_PCA	3.410/3.101/2.704	2.753/2.489/2.160	4.412/4.233/3.883	3.556/3.402/3.106
FM_PLS	0.798/0.654/0.623	0.646/0.525/0.498	1.107/0.693/0.634	0.892/0.556/0.507
$n = 100/200/400$	$\sigma = 0.3/0.9,\ P = 50$		$\sigma = 0.3/0.9,\ P = 70$	
MCP	**0.941/0.920/0.909**	**0.761/0.739/0.727**	**0.965/0.921/0.910**	**0.780/0.739/0.728**
E-SCAD	0.954/0.926/0.913	0.771/0.743/0.730	0.985/0.930/0.914	0.795/0.746/0.730
Autometrics	0.954/0.926/**0.909**	0.768/0.744/**0.727**	1.017/0.938/0.916	0.823/0.752/0.733
FM_PCA	3.478/3.176/2.791	2.809/2.549/2.230	4.467/4.281/3.941	3.601/3.440/3.153
FM_PLS	1.181/0.983/0.935	0.956/0.789/0.748	1.507/1.040/0.951	1.215/0.834/0.760

Note. Bold values indicate a better forecast.

extreme autocorrelation, E-SCAD outperformed the rival techniques under both small and moderate samples, but as we further augment the sample equal to 400, the MCP induced a more accurate forecast comparatively.

5. Real Data Analysis

After Monte Carlo experiments, this study performs real data analysis using big data. For real data analysis, we focus on two datasets: macroeconomic data and financial markets. In the context of both datasets, the study considers worker's remittances inflow and stock market data, respectively. It is a fact that many factors influence the worker's remittances inflow and the stock market. Among them, some covariates are recommended by economic and financial theories to be included in the model. Apart from this, a long list of variables has been recommended by past studies. This study considers all the possible determinants based on theories and literature as well to make a general model. In econometrics literature, such a model is known as the general unrestricted model (GUM).

5.1. Data Source. This study collects the annual data for Pakistan from 1973 to 2020. The data is sourced from the World Development Indicators (WDI), International

Table 3: Forecast comparison under autocorrelation from Monte Carlo simulation (Scenario 3).

Models	$\rho = 0.25,\ P = 50$		$\rho = 0.25,\ P = 70$	
$n = 100/200/400$	RMSE	MAE	RMSE	MAE
MCP	**1.167/1.078/1.056**	**0.943/0.866/0.845**	**1.254/1.110/1.065**	**1.012/0.892/0.851**
E-SCAD	1.175/1.091/1.062	0.952/0.877/0.850	**1.241**/1.124/1.074	**1.002**/0.904/0.859
Autometrics	1.192/1.100/1.064	0.963/0.884/0.851	1.392/1.126/1.071	1.121/0.908/0.858
FM_PCA	3.520/3.222/2.858	2.848/2.589/2.288	4.569/4.274/3.952	3.695/3.429/3.165
FM_PLS	1.568/1.231/1.119	1.268/0.990/0.896	1.972/1.367/1.166	1.591/1.101/0.932
$n = 100/200/400$	$\rho = 0.50,\ P = 50$		$\rho = 0.50,\ P = 70$	
MCP	1.324/**1.222/1.185**	1.073/**0.987/0.949**	1.448/**1.234/1.197**	1.177/**0.993/0.957**
E-SCAD	**1.318**/1.238/1.191	**1.068**/0.996/0.954	**1.382**/1.248/1.206	**1.122**/1.005/0.965
Autometrics	1.330/1.222/1.187	1.080/0.985/0.951	1.630/1.255/1.202	1.318/1.011/0.964
FM_PCA	3.570/3.279/2.916	2.889/2.624/2.333	4.607/4.247/4.021	3.716/3.381/3.219
FM_PLS	1.720/1.392/1.258	1.389/1.121/1.005	2.108/1.503/1.303	1.702/1.206/1.042
$n = 100/200/400$	$\rho = 0.90,\ P = 50$		$\rho = 0.90,\ P = 70$	
MCP	2.953/2.408/**2.364**	2.449/1.997/**1.936**	3.608/2.538/**2.368**	2.961/2.100/**1.940**
E-SCAD	**2.714/2.380**/2.366	**2.267/1.976**/1.937	**3.039/2.498**/2.370	**2.525/2.069**/1.941
Autometrics	3.250/2.480/2.358	2.693/2.049/1.930	4.273/2.594/2.394	3.494/2.146/1.957
FM_PCA	4.165/3.871/3.563	3.387/3.126/2.868	5.051/4.735/4.506	4.111/3.810/3.609
FM_PLS	2.941/2.579/2.476	2.439/2.122/2.020	3.341/2.796/2.544	2.749/2.293/2.072

Note. Bold values indicate a better forecast.

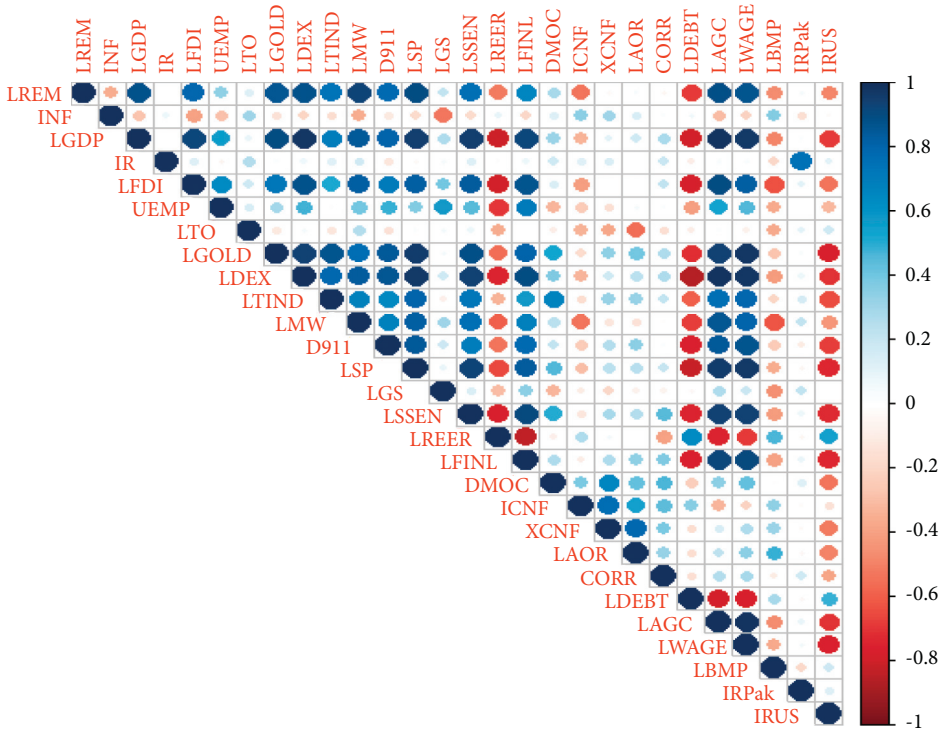

Figure 4: Pairwise correlation using macroeconomic data.

Financial Statistics (IFS), International Country Risk Guide, and State Bank of Pakistan. The few missing observations in the data set are replaced by averaging the neighbor observations. Most variables are transformed into logarithm form to ensure normality.

Details on the variables used for the analysis are given in Appendix Table 4.

5.2. Correlation Matrix. For empirical analysis, we split the data set into parts: observations from 1973 to 2007 are utilized to train the models and the remaining data are used to evaluate their forecasting performance. But before going to compute the forecast error, we discover the correlation structure among covariates through the visualization approach. In Figures 4 and 5, blue and red colors exhibit

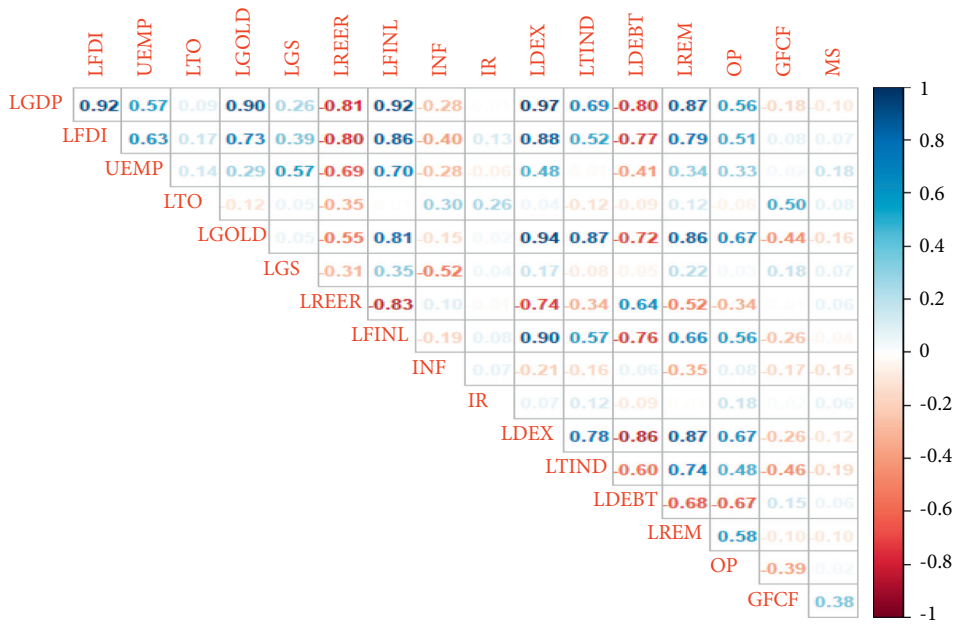

Figure 5: Pairwise correlation using financial data.

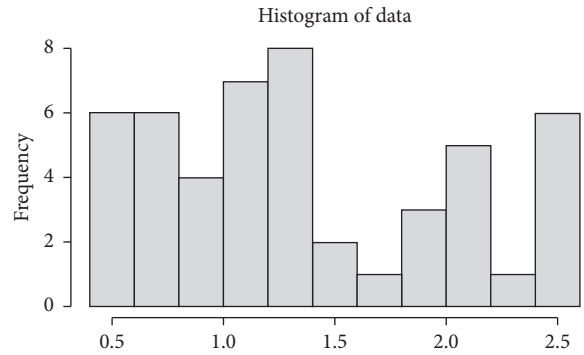

Figure 6: Histogram of Stock prices.

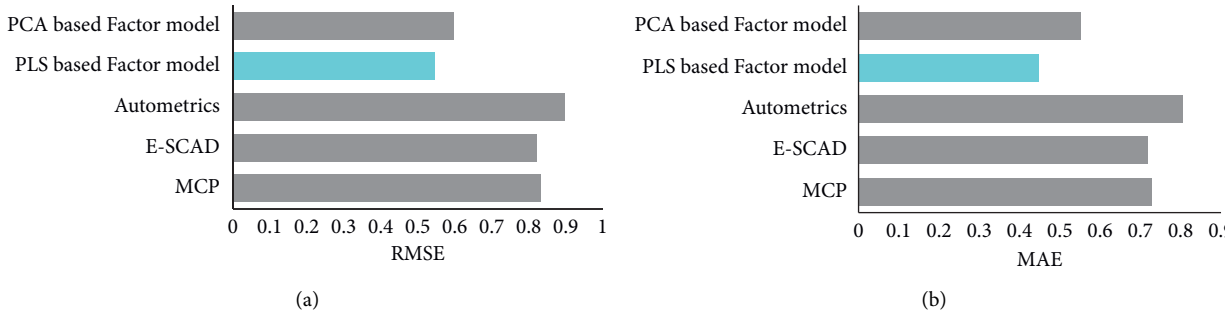

Figure 7: Forecast comparison using macroeconomic data.

positive and negative correlations, respectively. The colors' severity and the area of the circle are directly associated with correlation coefficients. On the right side of the correlogram, the legend color shows the correlation coefficients and the corresponding colors. We can observe that there are many dark color circles in blue and red, which clearly illustrate the high pairwise correlation. In other words, we can conclude that there exists high multicollinearity among predictors under both datasets. Figure 6 reveals that the distribution of stock market data is almost symmetric. Apart from this,

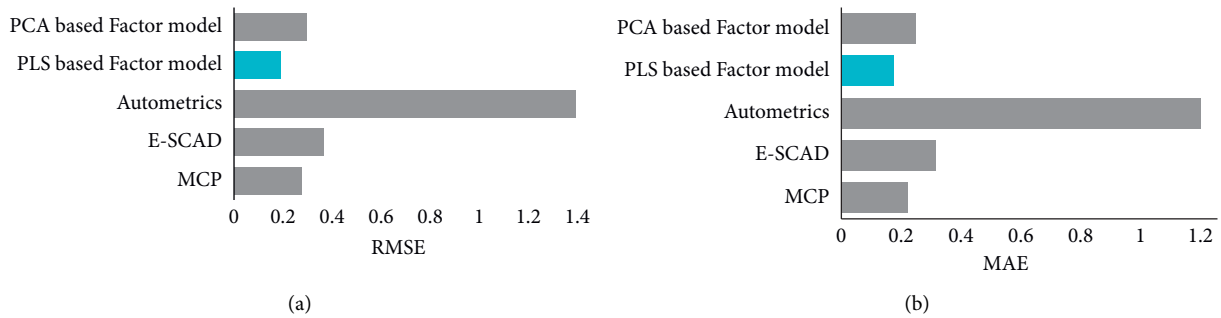

Figure 8: Forecast comparison using financial data.

Table 4: Methods' performance across various scenarios.

Scenarios	Three cases		
	Low	Medium	High
Multicollinearity	E-SCAD is best under a small sample. MCP is the best option in case of a large sample.	E-SCAD is best under a small sample. MCP is the best option in case of a large sample.	PLS-based factor model provides a better forecast under small sample. In case of a large sample, E-SCAD is superior.
Heteroscedasticity	MCP is best.	MCP is best.	MCP is best.
Autocorrelation	E-SCAD is best under a small sample. MCP is the best option using more data.	E-SCAD is best under a small sample. MCP is the best option using more data.	E-SCAD is best under a small sample. MCP is the best option using a large data set.

diagnostic tests revealed that the residuals of an estimated model are independently and identically distributed. As we have noted in simulation experiments that in presence of high multicollinearity, the PLS-based factor model outperformed the other methods in terms of forecast error particularly when the sample size is small. It reveals that PLS-based factor is more robust in such circumstances.

5.3. Forecast Comparison Based on Two Real Datasets. Root mean square error and mean absolute error are computed to ascertain the predictive ability of MCP, E-SCAD, Autometrics, and factor models based on PCA and PLS in Figures 7 and 8, respectively. The findings show that PLS-based factor model outperformed the rival methods in the out-of-sample forecast. It illustrates that PLS-based factor model has good predictive power than other competitor models, in terms of having the lowest forecast errors in multistep ahead forecast for the period (2008 to 2020). It supports the simulation results under both real datasets.

6. Concluding Remarks

This study compares factor models based on principal component analysis and partial least squares with classical approach (Autometrics) as well as shrinkage procedures (i.e., minimax concave penalty (MCP) and elastic smoothly clipped absolute deviation (E-SCAD)). The comparison is made under the presence of multicollinearity, heteroscedasticity, and autocorrelation with altering sample size and number of covariates. We carried out Monte Carlo experiments to compare all methods in terms of prediction. All methods are improving their performance with a growing sample size in all scenarios. Expanding the number of irrelevant and candidate variables negatively affects forecasting accuracy. In the presence of low and moderate multicollinearity, MCP often produced better forecasts comparatively except for the small number of observations, where E-SCAD is dominant. In the case of extreme multicollinearity, the PLS-based factor model is superior, but with increased sample sizes, the prediction accuracy of E-SCAD significantly boosts up as compared to other methods. In the presence of all schemes of heteroscedasticity, the performance of MCP is better than all competitor models. When the number of predictors is equal to 50, Autometrics provides a similar forecast as MCP in large samples. In the presence of low and moderate autocorrelation, the MCP showed an outstanding performance in terms of forecasting except for the small sample case where E-SCAD produced a remarkable forecast. In the case of extreme autocorrelation, E-SCAD outperformed the rival techniques under both the smallest and medium samples, but as we further augment the sample equal to 400, the MCP induced a more accurate forecast comparatively.

For empirical application, macroeconomic and financial datasets are used. To compare the forecasting performance of all methods, we divide the data into two parts (i.e., data over 1973–2007 as training data and data over 2008–2020 as testing data), using both datasets. All methods are trained on training data and subsequently, their performance was evaluated through testing data. Based on RMSE and MAE, the PLS-based factor model is more robust in terms of forecasting than competitor models. This study has several recommendations, reported in Table 4.

6.1. Limitations and Future Direction. The few limitations of this study are that it only focuses on linear models and has

considered yearly data. The simulation part of this study is restricted to Gaussian distributed errors, but in practice, this is not essential that the errors of a model are always normal. Hence, the research can be conducted to discover the forecasting performance of advanced statistical and machine learning techniques under nonnormal residuals as well as missing observations in the data set. This study can be expanded to examine the performance of nonlinear and nonparametric algorithms like artificial neural networks, random forests, support vector machines, etc.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this study.

Supplementary Materials

Appendix Table 4. Variables description. (*Supplementary Materials*)

References

[1] J. Bai and S. Ng, "Forecasting economic time series using targeted predictors," *Journal of Econometrics*, vol. 146, no. 2, pp. 304–317, 2008.

[2] B. S. Bernanke, J. Boivin, and P. Eliasz, "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach," *Quarterly Journal of Economics*, vol. 120, no. 1, pp. 387–422, 2005.

[3] D. E. Rapach, J. K. Strauss, and G. Zhou, "Out-of-sample equity premium prediction: combination forecasts and links to the real economy," *Review of Financial Studies*, vol. 23, no. 2, pp. 821–862, 2010.

[4] J. Li and W. Chen, "Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models," *International Journal of Forecasting*, vol. 30, no. 4, pp. 996–1015, 2014.

[5] H. R. Varian, "Big data: new tricks for econometrics," *The Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28, 2014.

[6] N. R. Swanson and W. Xiong, "Big data analytics in economics: what have we learned so far, and where should we go from here?" *Canadian Journal of Economics/Revue canadienne d'économique*, vol. 51, no. 3, pp. 695–746, 2018a.

[7] J. A. Doornik and D. F. Hendry, "Statistical model selection with big data," *Cogent Economics and Finance*, vol. 3, 2015.

[8] M. J. Artis, A. Banerjee, and M. Marcellino, "Factor forecasts for the UK," *Journal of Forecasting*, vol. 24, no. 4, pp. 279–298, 2005.

[9] J. Boivin and S. Ng, "Undertanding and comparing factor based forecasts," *International Journal of Central Banking*, vol. 1, no. 3, pp. 117–152, 2005.

[10] J. Boivin and S. Ng, "Are more data always better for factor analysis?" *Journal of Econometrics*, vol. 132, no. 1, pp. 169–194, 2006.

[11] M. Forni, M. Hallin, M. Lippi, and L. Reichlin, "The generalized dynamic factor model," *Journal of the American Statistical Association*, vol. 100, no. 471, pp. 830–840, 2005.

[12] N. A. Armah and N. R. Swanson, "Diffusion index models and index proxies: recent results and new direction," *European Journal of Pure and Applied Mathematics*, vol. 3, pp. 478–501, 2010.

[13] N. Ayi Armah and N. R. Swanson, "Seeing inside the black box: using diffusion index methodology to construct factor proxies in large scale macroeconomic time series environments," *Econometric Reviews*, vol. 29, no. 5-6, pp. 476–510, 2010.

[14] J. H. Stock and M. W. Watson, "Forecasting inflation," *Journal of Monetary Economics*, vol. 44, no. 2, pp. 293–335, 1999.

[15] J. H. Stock and M. W. Watson, "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1167–1179, 2002.

[16] J. H. Stock and M. W. Watson, "Implications of dynamic factor models for VAR analysis," *NBER Working Papers 11467*, National Bureau of Economic Research, Inc, Cambridge, MA, USA, 2005.

[17] J. H. Stock and M. W. Watson, "Chapter 10 forecasting with many predictors," in *Handbook of Economic Forecasting*, G. Elliott, C. Granger, and A. Timmermann, Eds., vol. 1, pp. 515–554, Elsevier, Chennai, India, 2006.

[18] J. H. Stock and M. W. Watson, "Generalized shrinkage methods for forecasting using many predictors," *Journal of Business and Economic Statistics*, vol. 30, no. 4, pp. 481–493, 2012.

[19] H. H. Kim and N. R. Swanson, "Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence," *Journal of Econometrics*, vol. 178, no. 2, pp. 352–367, 2014.

[20] H. H. Kim and N. R. Swanson, "Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods," *International Journal of Forecasting*, vol. 34, no. 2, pp. 339–354, 2018.

[21] J. L. Castle, M. P. Clements, and D. F. Hendry, "Forecasting by factors, by variables, by both or neither?" *Journal of Econometrics*, vol. 177, no. 2, pp. 305–319, 2013.

[22] J. L. Castle, J. A. Doornik, and D. F. Hendry, "Modelling non-stationary 'big data'," *International Journal of Forecasting*, 2020.

[23] M. Luciani, "Forecasting with approximate dynamic factor models: the role of non-pervasive shocks," *International Journal of Forecasting*, vol. 30, no. 1, pp. 20–29, 2014.

[24] J. T. Kristensen, "Diffusion indexes with sparse loadings," *Journal of Business & Economic Statistics*, vol. 35, no. 3, pp. 434–451, 2017.

[25] N. R. Swanson and W. Xiong, "Predicting interest rates using shrinkage methods, real-time diffusion indexes, and model combinations," *Working Paper*, Rutgers University, Cambridge, MA, USA, 2018.

[26] Y. Tu and T.-H. Lee, "Forecasting using supervised factor models," *Journal of Management Science and Engineering*, vol. 4, no. 1, pp. 12–27, 2019.

[27] N. R. Swanson, W. Xiong, and X. Yang, "Predicting interest rates using shrinkage methods, real-time diffusion indexes, and model combinations," *Journal of Applied Econometrics*, vol. 35, no. 5, pp. 587–613, 2020.

[28] K. Maehashi and M. Shintani, "Macroeconomic forecasting using factor models and machine learning: an application to

Japan," *Journal of the Japanese and International Economies*, vol. 58, Article ID 101104, 2020.
[29] H. Kim and K. Ko, "Improving forecast accuracy of financial vulnerability: PLS factor model approach," *Economic Modelling*, vol. 88, pp. 341–355, 2020.
[30] H. Kim, W. Shi, and H. H. Kim, "Forecasting financial stress indices in Korea: a factor model approach," *Empirical Economics*, vol. 59, no. 6, pp. 2859–2898, 2020.
[31] A. Abdić, E. Resić, and A. Abdić, "Modelling and forecasting GDP using factor model: an empirical study from Bosnia and Herzegovina," *Croatian Review of Economic, Business and Social Statistics*, vol. 6, no. 1, pp. 10–26, 2020.
[32] H. Kim and W. Shi, "Forecasting financial vulnerability in the USA: a factor model approach," *Journal of Forecasting*, vol. 40, no. 3, pp. 439–457, 2021.
[33] J. Bai and S. Ng, "Determining the number of factors in approximate factor models," *Econometrica*, vol. 70, no. 1, pp. 191–221, 2002.
[34] J. Bai and S. Ng, "Evaluating latent and observed factors in macroeconomics and finance," *Journal of Econometrics*, vol. 131, no. 1-2, pp. 507–537, 2006.
[35] A. Banerjee and M. Marcellino, *Factor-Augmented Error Correction Models*, CEPR Discussion Papers 6707, 2008.
[36] A. A. Ding and J. T. G. Hwang, "Prediction intervals, factor analysis models, and high-dimensional empirical linear prediction," *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 446–455, 1999.
[37] J.-M. Dufour and D. Stevanovic, "Factor-augmented VARMA models: identification, estimation, forecasting and impulse responses," *Working Paper*, McGill University, Montreal, Canada, 2010.
[38] S. Smeekes and E. Wijler, "Macroeconomic forecasting using penalized regression methods," *International Journal of Forecasting*, vol. 34, no. 3, pp. 408–430, 2018.
[39] J. Bai and S. Ng, "Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions," *Econometrica*, vol. 74, no. 4, pp. 1133–1150, 2006.
[40] H. Wold, "Soft modelling: the basic design and some extensions," *Systems under Indirect Observation, Part II*, North-Holland, Amsterdam, Netherlands, 1982.
[41] J. A. Doornik, "Econometric model selection with more variables than observations," *Working Paper*, Economics Department, University of Oxford, Oxford, UK, 2009.
[42] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
[43] L. Zeng and J. Xie, "Group variable selection via SCAD-L2," *Statistics*, vol. 48, no. 1, pp. 49–66, 2014.
[44] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.
[45] N. Xiao and Q.-S. Xu, "Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection," *Journal of Statistical Computation and Simulation*, vol. 85, no. 18, pp. 3755–3765, 2015.
[46] M. Qi and G. P. Zhang, "An investigation of model selection criteria for neural network time series forecasting," *European Journal of Operational Research*, vol. 132, no. 3, pp. 666–680, 2001.